On the first-order efficiency and asymptotic normality of maximum likelihood estimators obtained from dependent observations

Heijmans, R.D.H.; Magnus, J.R.

Published in:
Statistica Neerlandica

Publication date:
1986

ON THE FIRST-ORDER EFFICIENCY AND ASYMPTOTIC NORMALITY OF MAXIMUM LIKELIHOOD ESTIMATORS OBTAINED FROM DEPENDENT OBSERVATIONS

by R.D.H. Heijmans and J.R. Magnus

Abstract. In this paper we study the first-order efficiency and asymptotic normality of the maximum likelihood estimator obtained from dependent observations. Our conditions are weaker than usual, in that we do not require convergences in probability to be uniform or third-order derivatives to exist.

The paper builds on Witting and Nölle's result concerning the asymptotic normality of the maximum likelihood estimator obtained from independent and identically distributed observations, and on a martingale theorem by McLeish.

Key Words & Phrases: *Limiting distribution, dependent observations, vector martingales, maximum-likelihood.*

1 Introduction

1.1 Motivation

The key for establishing the efficiency and asymptotic normality of the maximum likelihood (ML) estimator seems to have been provided by DUGUÉ (1937). Later CRAMÉR ((1946), p. 500) used Dugué's method to provide a rigorous and much quoted proof for the case of independent and identically distributed (i.i.d.) observations. A huge literature on the efficiency and asymptotic normality of the ML estimator obtained from i.i.d. observations now exists, of which we mention in particular WITTING and NÖLLE'S ((1970), p. 78) approach, which goes back to BAHADUR (1967) and is the one adopted here.

When the observations are not i.i.d., the problem is more complex. SILVEY ((1961), p. 444) was among the first to investigate

> "... the extent to which the theory of the consistency and asymptotic normality of maximum-likelihood estimators, well established for observations on independent identically distributed random variables, carries over to more general stochastic processes. Everybody knows intuitively that it must do so to a considerable extent, but there seems to be a gap in the mathematical theory at this point and usable conditions which would enable one to establish these properties of maximum-likelihood estimation in particular cases are not distinguished by their multiplicity."

More than two decades later, this understatement has not yet lost its validity. The case where the observations are independent but not identically distributed was studied by BRADLEY and GART (1962), HOADLEY (1971), PHILIPPOU and ROUSSAS

(1973), and NORDBERG (1980). (See also AMEMIYA (1973), and GOURIEROUX and MONFORT (1981) for applications to the Tobit and logit models, respectively.) DIANANDA (1953), ANDERSON (1959), EICKER (1964), and SCHÖNFELD (1971) considered the case of m-dependent observations; BILLINGSLEY (1961) and ROUSSAS (1968) dealt with Markov processes which are stationary and ergodic; while ROZANOV ((1967), pp. 190-198) and HANNAN ((1970), pp. 220-229) studied stationary stochastic processes.

The literature for generally dependent observations includes WALD (1948), SILVEY (1961), BAR-SHALOM (1971), WEISS ((1971), (1973)), BHAT (1974), CROWDER (1976), BASAWA, FEIGIN and HEYDE (1976), and SWEETING (1980). While these papers are important contributions, the imposed conditions are generally either very restrictive or very difficult to verify.

In this article, which stands closest to the papers by CROWDER (1976) and BASAWA, FEIGIN and HEYDE (1976), a further attempt is made to establish intuitively appealing and verifiable conditions for the first-order efficiency and asymptotic normality of the ML estimator in a multi-parameter framework, assuming neither the independence nor the identical distribution of the observations. We believe that our conditions are weaker (and more readily applicable) than usual; in particular, we do not require convergences in probability to be uniform or third-order derivatives to exist.

1.2 Notation and set-up

The defining equality is denoted by $:=$, so that $x:=y$ defines x in terms of y. $\mathbb{N}:=\{1,2,..\}$, and $\mathbb{R}^p$ denotes the Euclidean space of dimension $p \geqslant 1$. To indicate the dimension of a vector, we often write $y_{(n)}:=(y_1,y_2,\cdots,y_n)$. Mathematical expectation, variance, and covariance are denoted by E, var, and cov, respectively. Measurable always means *Borel* measurable.

The set-up is as follows. Let $\{y_1,y_2,\cdots\}$ be a sequence of random variables, not necessarily independent or identically distributed. For each (fixed) $n \in \mathbb{N}$, let $y_{(n)}:=(y_1,y_2,\cdots,y_n)$ be defined on the probability space $(\mathbb{R}^n,\mathfrak{B}_n,\mathrm{P}_{n,\gamma})$ with values in $(\mathbb{R}^n,\mathfrak{B}_n)$, where $\mathfrak{B}_n$ denotes the minimal Borel field on $\mathbb{R}^n$. The following assumption will be made throughout.

Assumption 1.

For every (fixed) $n \in \mathbb{N}$ and $A \in \mathfrak{B}_n$,

$$\mathrm{P}_{n+1,\gamma}[A \times \mathbb{R}] = \mathrm{P}_{n,\gamma}[A].$$

This assumption, which relates the distribution functions of $y_{(n)}$ and $y_{(n+1)}$, is the consistency property used to prove Kolmogorov's Theorem (see RAO ((1973), p. 108)).

The joint density function of $y_{(n)}$ is a nonnegative and measurable function on $\mathbb{R}^n$, and is denoted by $h_n(\cdot\,;\gamma)$; it is defined with respect to μ_n, the Lebesgue measure on $(\mathbb{R}^n, \mathcal{B}_n)$. We assume that $h_n(\cdot\,;\gamma)$ is known, except for the values of a finite and fixed (i.e., not depending on n) number of parameters $\gamma := (\gamma_1, \gamma_2, \cdots, \gamma_p) \in \Gamma \subset \mathbb{R}^p$. The following two assumptions will also be made throughout.

Assumption 2.

For every (fixed) $n \in \mathbb{N}$, the probability measures $\{P_{n,\gamma}, \gamma \in \Gamma\}$ are absolutely continuous with respect to μ_n, i.e.,

$$P_{n,\gamma}[A] = \int_A h_n(y_{(n)};\gamma)\, d\mu_n(y_{(n)})$$

for every $A \in \mathcal{B}_n$.

Assumption 3.

For every (fixed) $n \in \mathbb{N}$, the set of positivity

$$\{y_{(n)} : h_n(y_{(n)};\gamma) > 0\}$$

is the same for all $\gamma \in \Gamma$.

Assumptions 2 and 3 are commonly made. Assumption 3 is important when the density is being differentiated. (Thus it is typically made when proving asymptotic normality, but not when proving consistency.) It implies however certain restrictions. For example, if $y_1, y_2, \cdots, y_n$ is a sample from the one-parameter distribution $f(y;\gamma) = \exp(\gamma - y)$, $y \geqslant \gamma$, or from the one-parameter uniform distribution $f(y;\gamma) = 1/\gamma$, $0 < y < \gamma$, then Assumption 3 is not satisfied. (In these two cases the ML estimator is consistent, but not asymptotically normal.) Assumptions 2 and 3 together imply that for every (fixed) $n \in \mathbb{N}$, the probability measures $\{P_{n,\gamma}, \gamma \in \Gamma\}$ are mutually absolutely continuous.

For every (fixed) $y \in \mathbb{R}^n$, the real-valued function

$$L_n(\gamma) := L_n(\gamma; y) := h_n(y;\gamma), \quad \gamma \in \Gamma, \tag{1}$$

is called the likelihood (function), and $\Lambda_n(\gamma) := \log L_n(\gamma)$ the loglikelihood (function). We let $L_0(\gamma) := 1$ and define the "conditional likelihood" (function)

$$g_n(\gamma) := g_n(\gamma; y) := \begin{cases} L_n(\gamma)/L_{n-1}(\gamma) & , \text{ if } L_{n-1}(\gamma) > 0 \\ 1 & , \text{ if } L_{n-1}(\gamma) = 0. \end{cases} \tag{2}$$

The true (but unknown) value of γ is denoted by γ_0, the interior of Γ by Γ^0, and we assume throughout that $\gamma_0 \in \Gamma^0$. *All probabilities and expectations are taken with respect to the true underlying distribution.* That is, we write P instead of P_{γ_0}, E instead of E_{γ_0}, etcetera.

For every (fixed) $y \in \mathbb{R}^n$, an ML *estimate* of γ_0 is a value $\hat{\gamma}_n(y) \in \Gamma$ with

$$L_n(\hat{\gamma}_n(y);y) = \sup_{\gamma \in \Gamma} L_n(\gamma;y). \tag{3}$$

Let M_n denote the set of $y \in \mathbb{R}^n$ for which an ML estimate exists, i.e.,

$$M_n := \bigcup_{\gamma \in \Gamma} \{y : y \in \mathbb{R}^n, L_n(\gamma;y) = \sup_{\phi \in \Gamma} L_n(\phi;y)\}. \tag{4}$$

If there exists, fore every $n \in \mathbb{N}$, a measurable function $\hat{\gamma}_n$ from $\mathbb{R}^n$ into Γ such that (3) holds for every $y \in M_n$, and a measurable subset M_n' of M_n such that $P(M_n') \rightarrow 1$ as $n \rightarrow \infty$, then we say that an ML estimator $\{\hat{\gamma}_n(y)\}$ of $\gamma_0 \in \Gamma$ exists *asymptotically almost surely*. (If Γ is open, compact, and interval, or, more generally, σ-compact, then M_n is itself measurable, see WITTING and NÖLLE (1970), p. 77). This notion of the existence of an ML estimator is somewhat less restrictive than usual.

In this paper we shall assume the existence (asymptotically almost surely) and consistency (weakly) of the ML estimator, and concentrate on proving first-order efficiency and asymptotic normality. Conditions for the existence and consistency of ML estimators are given in HEIJMANS and MAGNUS (1986a).

The theory developed in this paper assumes that the joint density of $(y_1, y_2, \cdots, y_n)$ is known (except, of course, for the value of the true parameter vector γ_0), but does not specify this function. In a sequel to this paper (HEIJMANS and MAGNUS (1986b)) we study the important special case of *normally distributed* (but generally dependent) observations, thereby demonstrating the applicability and strength of our results.

1.3 Outline of the paper

In section 2 we discuss first-order efficiency and prove Theorem 1, which generalizes WITTING and NÖLLE'S (1970) delicate result for the i.i.d. case to dependent observations. In section 3 we briefly review some results from the theory of vector martingales and give sufficient conditions for a vector martingale array to be asymptotically normally distributed. This result (Proposition 1) is a multivariate generalization of MCLEISH (1974). We put Theorem 1 and Proposition 1 together in section 4, where we prove asymptotic normality of the ML estimator (Theorem 2).

2 First-Order Efficiency

2.1 Background

Except in cases where the ML estimator admits of a closed form, a proof (as far as we are aware) of first-order efficiency of the ML estimator (RAO (1973), p. 348) typically proceeds as follows. Suppose for simplicity that only one parameter $\gamma_0 \in \Gamma$ is to be estimated, where Γ is an open interval on the real line, and let $\hat{\gamma}_n \in \Gamma$ be a weakly consistent ML estimator based on n observations. Developing $d\Lambda_n(\gamma)/d\gamma$ at $\hat{\gamma}_n$ in a first-order Taylor series about the true value γ_0, we obtain

$$0 = \frac{d\Lambda_n(\hat{\gamma}_n)}{d\gamma} = \frac{d\Lambda_n(\gamma_0)}{d\gamma} + \frac{d^2\Lambda_n(\tilde{\gamma}_n)}{d\gamma^2}(\hat{\gamma}_n - \gamma_0), \tag{5}$$

where

$$\tilde{\gamma}_n = t\gamma_0 + (1-t)\hat{\gamma}_n, \quad t \in (0,1).$$

This implies

$$h_n(\tilde{\gamma}_n)\left[\sqrt{n}(\hat{\gamma}_n - \gamma_0) - (1/\sigma_0^2)s_n(\gamma_0)\right] = -(1/\sigma_0^2)(h_n(\tilde{\gamma}_n) + \sigma_0^2)s_n(\gamma_0),$$

where

$$s_n(\gamma) := (1/\sqrt{n})d\Lambda_n(\gamma)/d\gamma,$$
$$h_n(\gamma) := (1/n)d^2\Lambda_n(\gamma)/d\gamma^2,$$

and $\sigma_0^2 > 0$ is arbitrary. Hence if we can demonstrate that any sequence of random variables $\{x_n\}$ which satisfies plim $x_n = 0$ as $n \to \infty$, also satisfies

$$\operatorname*{plim}_{n\to\infty} x_n s_n(\gamma_0) = 0, \tag{6}$$

and if we can also show that

$$\operatorname*{plim}_{n\to\infty} h_n(\gamma_0) = -\sigma_0^2, \tag{7}$$

and

$$\operatorname*{plim}_{n\to\infty} [h_n(\tilde{\gamma}_n) - h_n(\gamma_0)] = 0, \tag{8}$$

then it will follow that $\{\hat{\gamma}_n\}$ is first-order efficient, i.e.,

$$\operatorname*{plim}_{n\to\infty}[\sqrt{n}(\hat{\gamma}_n - \gamma_0) - (1/\sigma_0^2)s_n(\gamma_0)] = 0. \tag{9}$$

First-order efficiency thus states that $\hat{\gamma}_n$, appropriately centred and scaled, is asymptotically linearly related to the score function $s_n(\gamma_0)$.

Of these three conditions, (8) is the only difficult one. Consistency of $\hat{\gamma}_n$ is clearly not sufficient for (8) since Slutsky's theorem does not apply. One possibility to obtain sufficient conditions for (8) is to write

$$\begin{aligned}
&|h_n(\tilde{\gamma}_n) - h_n(\gamma_0)| \\
&= |[h_n(\tilde{\gamma}_n) + \sigma^2(\tilde{\gamma}_n)] - [h_n(\gamma_0) + \sigma_0^2] - [\sigma^2(\tilde{\gamma}_n) - \sigma_0^2]| \\
&\leqslant |h_n(\tilde{\gamma}_n) + \sigma^2(\tilde{\gamma}_n)| + |h_n(\gamma_0) + \sigma_0^2| + |\sigma^2(\tilde{\gamma}_n) - \sigma_0^2| \\
&\leqslant 2 \sup_{\gamma\in\Gamma} |h_n(\gamma) + \sigma^2(\gamma)| + |\sigma^2(\tilde{\gamma}_n) - \sigma_0^2|.
\end{aligned}$$

Hence, if $\sigma^2(\gamma)$ is a *continuous* function of γ with $\sigma^2(\gamma_0) = \sigma_0^2$, and if $h_n(\gamma)$ converges in probability to $-\sigma^2(\gamma)$ *uniformly* on Γ, i.e., $\operatorname{plim}_{n\to\infty} \sup_{\gamma\in\Gamma} |h_n(\gamma) + \sigma^2(\gamma)| = 0$, see BIERENS ((1981), p. 36), then (7) and (8) are both satisfied. This method was used among others by WILKS ((1962), pp. 104-105 and 360-362) and HALL and HEYDE ((1980), p. 160). Alternatively, if we can prove that $|d^3\Lambda_n(\gamma)/d\gamma^3| \leqslant H_n$, where H_n is independent of γ, and $\text{plim}(1/n)H_n < \infty$ as $n \to \infty$, then it follows that (8) is satisfied. This was CRAMÉR'S ((1946), p. 501) approach.

Conditions like uniform convergence in probability or the existence and uniform boundedness of a third-order derivative, are severe requirements, which are hard (in fact, often impossible) to verify. (See, for example, WHITE and DOMOWITZ ((1984), Theorem 2.3)). The purpose of this section is to provide sufficient (and verifiable) conditions for the first-order efficiency of the ML estimator which are weaker than usual. In particular, (i) it is not assumed that convergences in probability are uniform nor that third-order derivatives exist, and (ii) the observations are not required to be independent or identically distributed. We wish to demonstrate the following theorem which generalizes WITTING and NÖLLE'S ((1970), p. 78) result to dependent observations.

2.2 Theorem 1. First-Order Efficiency

Assume that

A1. an ML estimator $\{\hat{\gamma}_n\}$ exists asymptotically almost surely, and is weakly consistent;

A2. for every (fixed) $n \in \mathbb{N}$ and $y_{(n)} \in \mathbb{R}^n$, the likelihood $L_n(\gamma; y_{(n)})$ is twice continuously differentiable on Γ^0.

Let $l_n(\gamma) := \partial \Lambda_n(\gamma)/\partial\gamma$ (the score vector), and let $\{i_n\}$ be a sequence of nonrandom positive quantities with $i_n \to \infty$ as $n \to \infty$, such that

$$\lim_{M\to\infty} \limsup_{n\to\infty} \text{P}\left[(1/i_n) l'_n(\gamma_0) l_n(\gamma_0) > M\right] = 0. \tag{10}$$

(Such a sequence always exists; see point 5 in the discussion below.) Let $R_n(\gamma) := \partial^2 \Lambda_n(\gamma)/\partial\gamma\partial\gamma'$ (the Hessian matrix) with elements $R_{nij}(\gamma), i,j = 1, \cdots, p$, and assume that

A3. there exists a sequence of nonrandom positive quantities $\{j_n\}$ with $j_n \to \infty$ as $n \to \infty$, such that
   (i) $\underset{n\to\infty}{\text{plim}}(1/j_n) R_n(\gamma_0) = -G_0$, where G_0 is a positive definite $p \times p$ matrix, and
   (ii) for every $\epsilon > 0$ there exists a neighbourhood $N(\gamma_0)$ of γ_0 such that

$$\lim_{n\to\infty} \text{P}\left[(1/j_n) \sup_{\gamma \in N(\gamma_0)} |R_{nij}(\gamma) - R_{nij}(\gamma_0)| > \epsilon\right] = 0 \quad (i,j = 1, \cdots, p).$$

Then,

$$\underset{n\to\infty}{\text{plim}}\, i_n^{-\frac{1}{2}} (j_n(\hat{\gamma}_n - \gamma_0) - G_0^{-1} l_n(\gamma_0)) = 0. \tag{11}$$

Moreover, if $\{i_n\}$ can be chosen such that $\{i_n/j_n\}$ is bounded, then the sequence

$\{\hat{\gamma}_n\}$ is first-order efficient, i.e.,

$$\underset{n\to\infty}{\text{plim}}(j_n^{\frac{1}{2}}(\hat{\gamma}_n-\gamma_0)-j_n^{-\frac{1}{2}}G_0^{-1}l_n(\gamma_0))=0. \tag{12}$$

2.3 Discussion

Below are some remarks concerning Theorem 1.

1. It is essential that γ_0, the true parameter value, lies in the *interior* of Γ, because we require that the likelihood is differentiable at every point of some neighbourhood of γ_0. (MORAN (1971) studied the asymptotic behaviour of the ML estimator when γ_0 lies on the boundary of Γ.) Thus we can not permit that the set Γ is countable (as it would be, for instance, in a sample from the $\chi^2(r)$ distribution), since Γ would not, in that case, contain any interior point.
2. We assume that an ML estimator $\{\hat{\gamma}_n\}$ exists only *asymptotically* almost surely. That is, we allow for the possibility that, for finite n, the set of observations $(y_1, \cdots, y_n)$ which do not yield an ML estimate, has a positive probability.
3. We do not need an explicit identification condition. It is easy to see, however, that necessary for the consistency of $\{\hat{\gamma}_n\}$ is the condition

$$\liminf_{n\to\infty} P(\{y\in\mathbb{R}^n: L_n(\gamma;y)\neq L_n(\gamma_0;y)\})>0$$

   for every $\gamma\neq\gamma_0\in\Gamma$. Thus "asymptotic identification" in this sense is implicit in the consistency requirement.
4. Condition A2 requires that $L_n(\gamma;y)$ is twice continuously differentiable on Γ^0 for *every* $y\in\mathbb{R}^n$ (rather than for *almost every* $y\in\mathbb{R}^n$). The reason for this lies in the fact that we can not otherwise be certain that the expression in A3(ii),

$$\sup_{\gamma\in N(\gamma_0)} |R_{nij}(\gamma)-R_{nij}(\gamma_0)|,$$

   is measurable. See WITTING and NÖLLE ((1970), result A7.7, p. 185). We also need the continuity of the second derivatives of $L_n(\gamma;y)$ here.
5. A sequence $\{i_n\}$ satisfying (10) always exists. This follows because, in general, if x is a random variable then for every $\epsilon>0$ there exists a $\delta=\delta(\epsilon)>0$ such that $P(x>\delta)<\epsilon$. In particular, there exists a sequence $\{k_n\}$ such that

$$P(l'_n(\gamma_0)l_n(\gamma_0)>k_n)<1/n.$$

   Thus, choosing $i_n:=k_n/M$, we obtain

$$P\left[(1/i_n)l'_n(\gamma_0)l_n(\gamma_0)>M\right]\to 0 \text{ as } n\to\infty.$$

6. In the classical case we have $i_n=j_n=n$, but in general the two sequences $\{i_n\}$ and $\{j_n\}$ may depend on γ_0. The smaller we are able to choose $\{i_n\}$ the more powerful is conclusion (11). An obvious choice is $i_n:=El'_n(\gamma_0)l_n(\gamma_0)$, provided the expectation exists and $i_n\to\infty$ as $n\to\infty$. Then (10) holds (see Lemma A.3(ii)

in the appendix), and $\{\hat{\gamma}_n\}$ is first-order efficient if Al-A3 are satisfied and

$$\{\mathrm{E}\; l'_n(\gamma_0) l_n(\gamma_0)/j_n\} \tag{13}$$

is bounded in n. In fact (13) can be somewhat weakened; it suffices that

$$\{\mathrm{E}\left[l'_n(\gamma_0) l_n(\gamma_0)\right]^r / j_n^r\} \tag{14}$$

is bounded in n for some $r>0$. (To see this, take

$$i_n := \left[\mathrm{E}\left[l'_n(\gamma_0) l_n(\gamma_0)\right]^r\right]^{1/r},$$

and note that i_n is a nondecreasing function of r for $r>0$, see RAO ((1973), 2.1(b), p.143).)

7. Condition A3(ii) is similar to condition (4.4) of CROWDER ((1976) p.48) and condition (A3)(iii) of BASAWA, FEIGIN and HEYDE ((1976) p.267). In case the observations $(y_1, \cdots, yy_n)$ are i.i.d. with a common denisty $f(y;\gamma)$, it is not difficult to verify that A3(ii) holds if

$$\mathrm{E} \sup_{\gamma \in N(\gamma_0)} \left| \frac{\partial^2 \log(f(y;\gamma)}{\partial\gamma_i \partial\gamma_j} - \frac{\partial^2 \log(f(y;\gamma_0)}{\partial\gamma_i \partial\gamma_j} \right| < \infty$$

for $i,j=1, \cdots, p$ and some neighbourhood $N(\gamma_0)$ of γ_0.

8. If the observations are not i.i.d., verification of A3(ii) is more difficult, but still possible. (PAGAN (1980), HEIJMANS and MAGNUS (1986b)) In any case, condition A3(ii) is a fairly weak condition, compared to any assumption on uniform convergence in probability or the uniform boundedness of third-order derivatives. In fact, it can be shown that if $G(\gamma)=(G_{ij}(\gamma))$ is a *continuous* $p \times p$ matrix function of γ with $G(\gamma_0)=G_0$, and if A3(i) holds, then a necessary and sufficient condition for A3(ii) is that for every $\epsilon>0$ there exists a neighbourhood $N(\gamma_0)$ such that

$$\lim_{n\to\infty} \mathrm{P}[\sup_{\gamma \in N(\gamma_0)} |(1/j_n) R_{nij}(\gamma) + G_{ij}(\gamma)| > \epsilon| = 0, \tag{15}$$

$(i,j=1, \cdots, p)$. The neighbourhood $N(\gamma_0)$ in (15) may depend on ϵ. If we require that one neighbourhood $N(\gamma_0)$ serves equally well for *every* $\epsilon>0$, then we obtain the much stronger condition

$$\operatorname*{plim}_{n\to\infty} \sup_{\gamma \in N(\gamma_0)} |(1/j_n) R_{nij}(\gamma) + G_{ij}(\gamma)| = 0, \tag{16}$$

$(i,j=1, \cdots, p)$, which amounts to convergence in probability of $(1/j_n)R_n(\gamma)$ to $-G(\gamma)$ uniformly on $N(\gamma_0)$. Hence our assumption A3(ii) is much weaker than the usual assumption (16).

9. It is also true that A3(ii) is weaker than the alternative assumption of uniformly bounded third-order derivatives. To be precise, if we assume that for $i,j,h=1, \cdots, p$

$$\left|\frac{\partial^3 \Lambda_n(\gamma)}{\partial\gamma_i \partial\gamma_j \partial\gamma_h}\right| \leqslant H_{nijh}$$

for all $\gamma \in \Gamma^0$ (or at least for all γ in some neighbourhood of γ_0), where H_{nijh} is independent of γ, and also that

$$\operatorname*{plim}_{n\to\infty}(1/j_n)H_{nijh} < \infty,$$

then we can easily show (using the mean-value theorem for random functions (JENNRICH (1969), Lemma 3)), that A3(ii) holds.

2.4 Proof of Theorem 1

Let M_n be the set of y in $\mathbb{R}^n$ for which an ML estimates exists, i.e.,

$$M_n := \bigcup_{\gamma\in\Gamma}\{y : y\in\mathbb{R}^n, L_n(\gamma;y) = \sup_{\phi\in\Gamma} L_n(\phi;y)\}.$$

Since, by A1, an ML estsimator $\{\hat{\gamma}_n\}$ exists asymptotically almost surely, there exists a sequence $\{M_n'\}$ such that M_n' is a measurable subset of M_n for every $n\in\mathbb{N}$, and $P(M_n')\to 1$ as $n\to\infty$. (See section 1.2.)

Let $N_0(\gamma_0)$ be a neighbourhood of γ_0 such that its closure is a subset of Γ^0, and define

$$V_n := \{y : y\in\mathbb{R}^n, \hat{\gamma}_n \in N_0(\gamma_0)\}.$$

Also define a new random variable $\overline{\gamma}_n$ as

$$\overline{\gamma}_n := \begin{cases} \hat{\gamma}_n, & \text{if } y\in M_n'\cap V_n; \\ \gamma_0, & \textit{otherwise.} \end{cases}$$

Condition A2 together with Jennrich's mean-value theorem for random functions (JENNRICH (1969)) Lemma 3)) allows us to develop the $h-th$ partial derivative of $\Lambda_n(\gamma)$ at $\overline{\gamma}_n$ in a first-order Taylor series about γ_0,

$$\frac{\partial\Lambda_n(\overline{\gamma}_n)}{\partial\gamma_h} = \frac{\partial\Lambda_n(\gamma_0)}{\partial\gamma_h} + e'_h R_n(\tilde{\gamma}_{nh})(\overline{\gamma}_n - \gamma_0), \quad (h=1,\cdots,p), \tag{17}$$

where

$$\tilde{\gamma}_{nh} := \begin{cases} t_{nh}\gamma_0 + (1-t_{nh})\hat{\gamma}_n, & t_{nh}\in(0,1), \ \text{if } y\in M_n'\cap V_n; \\ \gamma_0, & \textit{otherwise,} \end{cases}$$

and e_h is a $p\times 1$ vector with 1 in its $h-th$ position and zeros elsewhere. (We consider the scalar function $\partial\Lambda_n(\gamma)/\partial\gamma_h$ rather than the vector function $\partial\Lambda_n(\gamma)/\partial\gamma$, because the mean-value theorem is not valid for vector functions, see APOSTOL

((1974), p. 355).) Note that t_{nh} itself is a random variable. Also note that t_{nh}, and hence $\tilde{\gamma}_{nh}$, differs for each h.

From (17) we obtain

$$
\begin{aligned}
&(1/j_n)e'_h R_n(\tilde{\gamma}_{nh})(k_n(\hat{\gamma}_n-\gamma_0)-G_0^{-1} i_n^{-\frac{1}{2}} l_n(\gamma_0)) \\
&=e'_h \Big[(1/j_n)R_n(\tilde{\gamma}_{nh})k_n(\hat{\gamma}_n-\bar{\gamma}_n)+i_n^{-\frac{1}{2}} l_n(\bar{\gamma}_n) \\
&-[(1/j_n)R_n(\tilde{\gamma}_{nh})+G_0]G_0^{-1} i_n^{-\frac{1}{2}} l_n(\gamma_0)\Big] \qquad (18)
\end{aligned}
$$

where $k_n := i_n^{-\frac{1}{2}} j_n$. Now,

$$P(k_n(\hat{\gamma}_n-\bar{\gamma}_n)=0) \geqslant P(y \in M'_n \cap V_n),$$

and since $\{\hat{\gamma}_n\}$ exists asymptotically almost surely and is weakly consistent, the latter probability tends to 1, as $n \to \infty$. Hence,

$$\operatorname*{plim}_{n\to\infty} k_n(\hat{\gamma}_n-\bar{\gamma}_n)=0. \qquad (19)$$

Similarly, the fact that

$$P(i_n^{-\frac{1}{2}} l_n(\bar{\gamma}_n)=0) \geqslant P(y \in M'_n \cap V_n),$$

implies that

$$\operatorname*{plim}_{n\geqslant\infty} i_n^{-\frac{1}{2}} l_n(\bar{\gamma}_n)=0. \qquad (20)$$

Next, let us write

$$R_n(\tilde{\gamma}_{nh})=R_n(\gamma_0)+(R_n(\tilde{\gamma}_{nh})-R_n(\gamma_0)).$$

Now, $(1/j_n)R_n(\gamma_0)$ converges to $-G_0$ as $n \to \infty$ (condition A3(i)). Further, let $\epsilon>0$ be arbitrary and choose a neighbourhood $N_1(\gamma_0) \subset N_0(\gamma_0)$ of γ_0 such that A3(ii) holds. (The choice of $N_1(\gamma_0)$ may depend on ϵ.) Then

$$
\begin{aligned}
&P\Big[(1/j_n)|R_{nij}(\tilde{\gamma}_{nh})-R_{nij}(\gamma_0)|>\epsilon\Big] \\
&\leqslant P\Big[(1/j_n)|R_{nij}(\tilde{\gamma}_{nh})-R_{nij}(\gamma_0)|>\epsilon \text{ and } \tilde{\gamma}_{nh} \in N_1(\gamma_0)\Big]+P(\tilde{\gamma}_{nh} \notin N_1(\gamma_0)) \\
&\leqslant P\Big[(1/j_n) \sup_{\gamma \in N_1(\gamma_0)} |R_{nij}(\gamma)-R_{nij}(\gamma_0)|>\epsilon\Big]+P(\tilde{\gamma}_{nh} \notin N_1(\gamma_0)),
\end{aligned}
$$

which tends to 0 as $n \to \infty$, because of condition A3(ii) and the weak consistency of $\{\hat{\gamma}_n\}$. Since ϵ was chosen arbitrarily, we obtain $\operatorname{plim}(1/j_n)(R_n(\tilde{\gamma}_{nh})-R_n(\gamma_0))=0$ as $n \to \infty$, and thus

$$\operatorname*{plim}_{n\to\infty}(1/j_n)R_n(\tilde{\gamma}_{nh})=-G_0, \quad (h=1, \cdots, p). \qquad (21)$$

From (21) and the fact that the sequence $\{i_n\}$ satisfies (10), by construction, we obtain

$$\operatorname*{plim}_{n\to\infty}\left[(1/j_n)R_n(\tilde{\gamma}_{nh})+G_0\right]G_0^{-1}i_n^{-\frac{1}{2}}l_n(\gamma_0)=0, \tag{22}$$

using Lemma A.2 in the appendix. From (19) — (22) then follows that the right-hand side, and hence the left hand side, of (18) tends to zero in probability, i.e.

$$\operatorname*{plim}_{n\to\infty}(1/j_n)e'_hR_n(\tilde{\gamma}_{nh})(k_n(\hat{\gamma}_n-\gamma_0)-G_0^{-1}i_n^{-\frac{1}{2}}l_n(\gamma_0)) = 0. \tag{23}$$

Let us define the $p\times p$ matrix (not, in general, symmetric)

$$\tilde{R}_n := (R_n(\tilde{\gamma}_{n1})e_1, \cdots, R_n(\tilde{\gamma}_{np})e_p).$$

Then, from (23),

$$\operatorname*{plim}_{n\to\infty}(1/j_n)\tilde{R}'_n\left[k_n(\hat{\gamma}_n-\gamma_0)-G_0^{-1}i_n^{-\frac{1}{2}}l_n(\gamma_0)\right] = 0.$$

But (21) implies that $(1/j_n)\tilde{R}_n$ converges in probability to a *non-singular* matrix $-G_0$. Hence, by Lemma A.1(ii),

$$\operatorname*{plim}_{n\to\infty}\left[k_n(\hat{\gamma}_n-\gamma_0)-G_0^{-1}i_n^{-\frac{1}{2}}l_n(\gamma_0)\right] = 0.$$

This proves (11). (Recall that $k_n=i_n^{-\frac{1}{2}}j_n$.) If $\{i_n/j_n\}$ is bounded, then (12) is an immediate consequence of (11) and Lemma A.1(iii).

This concludes the proof of Theorem 1.

3 An Interlude: Vector Martingales

In this section we first present a short review of vector martingales, and then generalize a theorem of McLeish, thus obtaining sufficient conditions for the asymptotic normality of a vector martingale array.

Let $(\Omega,\mathbb{F},\mathrm{P})$ be a probability space: Ω is a set, $\mathbb{F}$ a σ-field of subsets of Ω, and P is a probability measure defined on $\mathbb{F}$.

Definition 3.1

Let $\{s_n, n\in\mathbb{N}\}$ be a sequence of random vectors (of fixed dimension) on Ω, and $\{\mathbb{F}_n, n\in\mathbb{N}\}$ a sequence of sub σ-fields of $\mathbb{F}$ generated by $s_1, \cdots, s_n$. If

(i) $\mathrm{E}|s_n|<\infty$ and

(ii) $\mathrm{E}(s_n|\mathbb{F}_{n-1})=s_{n-1}$ a.s. for all $n\geqslant 2$,

then the sequence $\{s_n, n\in\mathbb{N}\}$ is said to be a *vector martingale.*

Definition 3.2.

Let $\{s_n, n\in\mathbb{N}\}$ be a vector martingale. Then the sequence of random vectors $\{x_n, n\in\mathbb{N}\}$ with

$$x_1 := s_1 \quad\text{and}\quad x_n := s_n-s_{n-1}, \quad n\geqslant 2$$

is called a *vector martingale difference* (VMD).

Definition 3.3.
If $\{x_{n,t},\ 1 \leqslant t \leqslant n\}$ is a vector martingale difference for each $n \in \mathbb{N}$, then the double sequence $\{x_{n,t},\ 1 \leqslant t \leqslant n,\ n \in \mathbb{N}\}$ is called a *vector martingale difference array* (VMDA).

If $\{s_n\}$ is a sequence of random variables, rather than of random vectors, we delete the word "vector" and call it simply a martingale. Similarly, we call $\{x_n\}$ a martingale difference and $\{x_{n,t}\}$ a martingale difference array in the scalar case.

The following five properties are immediate consequences of the definitions:

$$E(s_n | s_1, \cdots, s_{n-1}) = s_{n-1} \quad \text{a.s. } (n \geqslant 2),$$

$$E(x_n | s_1, \cdots, s_{n-1}) = E(x_n | x_1, \cdots, x_{n-1}) = 0 \quad \text{a.s. } (n \geqslant 2),$$

$$E(x_{n,t} | x_{n,1}, \cdots, x_{n,t-1}) = 0 \quad \text{a.s. } (2 \leqslant t \leqslant n),$$

$$Es_n = Ex_1 \quad (n \in \mathbb{N}),$$

$$Ex_n = 0 \quad (n \geqslant 2).$$

If a sequence of random variables $\{x_n, n \in \mathbb{N}\}$ is a martingale difference, and $Ex_n^2 < \infty$ then

$$\operatorname{cov}(x_n, x_t) = 0 \quad (t \neq n \in \mathbb{N}).^{\bullet}$$

Also, if $x_1, x_2, \cdots$ are stochastically independent, and $Ex_1 < \infty, Ex_n = 0$ $(n \geqslant 2)$, then

$$E(x_n | x_1, \cdots, x_{n-1}) = Ex_n = 0 \quad (n \geqslant 2)$$

and hence $\{x_n, n \in \mathbb{N}\}$ is a martingale difference. This shows that a square-integrable (vector) martingale difference is intermediate between stochastic independence and uncorrelatedness. Since independence is not a vital requirement for the key limit theorems of probability — it is sufficient to consider random variables which form a martingale difference —, and since most stochastic processes can be easily transformed into martingale differences, † martingale limit theory has become an important tool in modern probability theory.

While a vector martingale is a straightforward generalization of a martingale, it is not true that a vector of martingales is necessarily a vector martingale. For example, let $\{s_n, n \in \mathbb{N}\}$ be a martingale and define $t_n := (s_n, s_{n-1})', n \geqslant 2$, and $t_1 := (s_1, 0)'$.

• In fact a stronger result is true: if $\{x_n\}$ is a martingale difference and $Ex_n^2 < \infty$, then $\operatorname{cov}(x_n, \psi(x_1, \cdots, x_{n-1})) = 0, n \in \mathbb{N}$, for every scalar function $\psi(.)$ which is bounded and measurable. See also Doob (1953, p. 92).

† Let $\{x_n, n \in \mathbb{N}\}$ be a sequence of random variables, with $E|x_n| < \infty$, and define $y_n := x_n - E(x_n | x_1, \cdots, x_{n-1}), n \geqslant 2$, and $y_1 := x_1 - Ex_1$. Then $\{y_n, n \in \mathbb{N}\}$ is a martingale difference.

Then $\{t_n, n \in \mathbb{N}\}$ is not a vector martingale (in fact, not even a sequence of uncorrelated vectors), since $E(t_n | t_1, \cdots, t_{n-1}) = (s_{n-1}, s_{n-1})' \neq t_{n-1}$. The following lemma gives sufficient conditions for a vector of martingales to be a vector martingale.

Lemma 1.
Consider p sequences of random variables $\{s_n^{(i)}, n \in \mathbb{N}\}$, $i = 1, \cdots, p$. Suppose there exists a sequence of random variables $\{y_n, n \in \mathbb{N}\}$ such that for $i = 1, \cdots, p$ and all $n \in \mathbb{N}$, $s_n^{(i)}$ is a measurable function of $y_1, \cdots, y_n$ with $E|s_n^{(i)}| < \infty$, and

$$E(s_n^{(i)} | y_1, \cdots, y_{n-1}) = s_{n-1}^{(i)}.$$

Then the sequence of random vectors $\{s_n, n \in \mathbb{N}\}$ with $s_n := (s_n^{(1)}, \cdots, s_n^{(p)})$, is a vector martingale.

Proof.
This is obvious, because

$$E(s_n | s_1, \cdots, s_{n-1}) = E(s_n | y_1, \cdots, y_{n-1}) = s_{n-1}. \|$$

The following crucial result gives a vector generalization of a result of McLEISH (1974), establishing sufficient conditions for a vector martingale array to be asymptotically normally distributed.

Proposition 1.
Let $\{x_{n,t},\ 1 \leqslant t \leqslant n, n \in \mathbb{N}\}$ be a vector martingale difference array satisfying
(i) $E(\max_{1 \leqslant t \leqslant n} x'_{n,t} x_{n,t})$ is bounded in n,
(ii) $\text{plim}(\max_{1 \leqslant t \leqslant n} x'_{n,t} x_{n,t}) = 0$, as $n \to \infty$, and
(iii) $\text{plim}(\Sigma_{t=1}^n x_{n,t} x'_{n,t}) = K$, as $n \to \infty$,
where K is a finite positive semidefinite matrix. Then

$$s_n := \sum_{t=1}^{n} x_{n,t} \xrightarrow{L} N(0, K).$$

Proof.
Let c be an arbitrary $p \times 1$ vector of real constants (where p is the dimension of the random vector $x_{n,t}$, and let $v_{n,t} := c' x_{n,t}$. Then $\{v_{n,t},\ 1 \leqslant t \leqslant n,\ n \in \mathbb{N}\}$ is a (univariate) martingale difference array. Conditions (i)-(iii) on $\{x_{n,t}\}$ imply that $\{v_{n,t}\}$ satisfies: (a) $E \max_{1 \leqslant t \leqslant n} v_{n,t}^2$ is bounded in n; (b) $\text{plim} \max_{1 \leqslant t \leqslant n} v_{n,t}^2 = 0$ as $n \to \infty$; and (c) $\text{plim}\ \Sigma_{t=1}^n v_{n,t}^2 = c' K c$ as $n \to \infty$. (Use Cauchy-Schwarz' inequality.) It then follows from HALL and HEYDE ((1980), Theorem 3.2) that

$$c' s_n = \sum_{t=1}^{n} v_{n,t} \xrightarrow{L} N(0, c' K c).$$

This holds for any real vector c. Hence $s_n \xrightarrow{L} N(0, K)$, see (RAO (1973), 2c.(xi), p. 123). ||

For a discussion of the relationship between conditions (i) and (ii) and the Lindeberg condition, the reader is referred to McLeish (1974), pp. 621-622) and Hall and Heyde ((1980), section 3.2).

4 Asymptotic normality

4.1 Background

Theorem 1 gives conditions under which the ML estimator $\hat{\gamma}_n$, appropriately centred and scaled, is asymptotically linearly related to the score function $s_n(\gamma_0):=(1/\sqrt{n})\partial\Lambda_n(\gamma_0)/\partial\gamma$. Thus, if we can show that $\{s_n(\gamma_0)\}$ is asymptotically normally distributed, then $\hat{\gamma}_n$ too will be asymptotically normally distributed.

Since the sequence $\{\partial\Lambda_n(\gamma_0)/\partial\gamma\}$ is usually a (vector) martingale (Silvey (1961), p. 450), the sequence $\{s_n(\gamma_0)\}$ is usually a (vector) martingale array. Proposition 1 then enables us to prove the asymptotic normality of $\{s_n(\gamma_0)\}$.

Combining Theorem 1 and Proposition 1, we can now state our main result, which gives the asymptotic normality of the ML estimator under conditions which we believe are both general and readily verifiable.

4.2 Theorem 2: Asymptotic Normality

Assume that

B1. an ML estimator $\{\hat{\gamma}_n\}$ exists asymptotically almost surely, and is weakly consistent;
B2. for every (fixed) $n\in\mathbb{N}$ and $y_{(n)}\in\mathbb{R}^n$, the likelihood $L_n(\gamma;y_{(n)})$ is twice continuously differentiable on Γ^0.

Let $l_n:=\partial\Lambda_n(\gamma_0)/\partial\gamma$ (the $p\times 1$ score vector evaluated at γ_0), $\xi_n:=\partial\log g_n(\gamma_0)/\partial\gamma$ (a $p\times 1$ vector with components ξ_{nj}, $j=1,\cdots,p$), and $R_n(\gamma):=\partial^2\Lambda_n(\gamma)/\partial\gamma\partial\gamma'$ (the $p\times p$ Hessian matrix with elements $R_{nij}(\gamma)$, $i,j=1,\cdots,p$), and assume that

B3. $\mathrm{E}\xi_{nj}^4<\infty$, $\quad(n\in\mathbb{N},j=1,\cdots,p)$;
B4. $\mathrm{E}(\xi_{nj}|y_1,\cdots,y_{n-1})=0$ a.s. $(n\geqslant 2,j=1,\cdots,p)$;
B5. $i_n:=\mathrm{E}l'_nl_n\to\infty$ as $n\to\infty$;
B6. there exists a finite positive semidefinite $p\times p$ matrix K_0 such that $\lim_{n\to\infty}(1/i_n)\mathrm{E}l_nl'_n=K_0$;
B7. $\mathrm{plim}_{n\to\infty}(1/i_n)\max_{1\leqslant t\leqslant n}\xi_{tj}^2=0$, $(j=1,\cdots,p)$;
B8. $\lim_{n\to\infty}(1/i_n^2)\mathrm{var}\Sigma_{t=1}^n\mathrm{E}(\xi_{ti}\xi_{tj}|y_1,\cdots,y_{t-1})=0$, $(i,j=1,\cdots,p)$;
B9. $\lim_{n\to\infty}(1/i_n^2)\Sigma_{t=1}^n\mathrm{var}(\xi_{ti}\xi_{tj}-\mathrm{E}(\xi_{ti}\xi_{tj}|y_1,\cdots,y_{t-1}))=0$, $(i,j=1,\cdots,p)$;
B10. there exists a sequence of nonrandom positive quantities $\{j_n\}$ with $j_n\to\infty$ as $n\to\infty$, such that $\mathrm{plim}_{n\to\infty}(1/j_n)\mathrm{tr}R_n(\gamma_0)=-1$;
B11. there exists a finite positive definite $p\times p$ matrix G_0 such that

$$\operatorname*{plim}_{n\to\infty}(1/j_n)R_n(\gamma_0)=-G_0;$$

B12. for every $\epsilon>0$ there exists a neighbourhood $N(\gamma_0)$ of γ_0 such that

$$\lim_{n\to\infty} \mathrm{P}\left[(1/j_n) \sup_{\gamma\in N(\gamma_0)} |R_{nij}(\gamma)-R_{nij}(\gamma_0)|>\epsilon\right]=0, \quad (i,j=1,\cdots,p).$$

Then the sequence $\{\hat{\gamma}_n\}$ is asymptotically normally distributed, i.e.,

$$i_n^{-\frac{1}{2}} j_n(\hat{\gamma}_n-\gamma_0)\xrightarrow{L} \mathrm{N}(0,G_0^{-1}K_0G_0^{-1}).$$

Moreover, if $\{i_n/j_n\}$ is bounded, then $\{\hat{\gamma}_n\}$ is first-order efficient, i.e.,

$$\operatorname*{plim}_{n\to\infty}(j_n^{\frac{1}{2}}(\hat{\gamma}_n-\gamma_0)-j_n^{-\frac{1}{2}}G_0^{-1}l_n)=0.$$

4.3 Discussion

1. Condition B4 implies (via Lemma 1) that $\{\partial\Lambda_n(\gamma_0)/\partial\gamma\}$ is a vector martingale. If we know, in addition, that the expectation of $\partial\Lambda_1(\gamma_0)/\partial\gamma$ vanishes, then we have first-order regularity, i.e.,

$$\mathrm{E}\partial\Lambda_n(\gamma_0)/\partial\gamma=0, \quad (n\in\mathbb{N}).$$

2. Condition B4 also ensures that the random variables in B9,

$$\{\xi_{ti}\xi_{tj}-\mathrm{E}(\xi_{ti}\xi_{tj}|y_1,\ldots,y_{t-1})\}_{t=1}^{n},$$

are uncorrelated (see footnote *), so that the variance of their sum equals the sum of their variances.

3. HALL and HEYDE ((1980), p. 157) stress the importance of the $p\times p$ "conditional information matrix" $F_n(\gamma_0)$ whose typical element is

$$F_{nij}(\gamma_0):=\sum_{t=1}^{n}\mathrm{E}(\xi_{ti}\xi_{tj}|y_1,\cdots,y_{t-1}).$$

In B5 we require that $i_n=\mathrm{tr}\mathrm{E}F_n(\gamma_0)\to\infty$, in B6 that $(1/i_n)\mathrm{E}F_n(\gamma_0)\to K_0$, and in B8 that $(1/i_n^2)\mathrm{var}F_{nij}(\gamma_0)\to 0$. Thus, we also have

$$\operatorname*{plim}_{n\to\infty}(1/i_n)F_n(\gamma_0)=K_0.$$

4. Conditions B8 and B9 together imply

$$\lim_{n\to\infty}(1/i_n^2)\mathrm{var}\sum_{t=1}^{n}\xi_{ti}\xi_{tj}=0, \quad (i,j=1,\cdots,p),$$

which is all that we need. The reason for having two conditions rather than one, is the belief that verifying B8 and B9 is the simplest way of establishing the above condition.

5. B5 and B6 imply $\mathrm{tr}K_0=1$; similarly B10 and B11 imply $\mathrm{tr}G_0=1$. In practice we may often choose $i_n=j_n$ and $K_0=G_0$, in which case

$$j_n^{\frac{1}{2}}(\hat{\gamma}_n-\gamma_0)\xrightarrow{L} \mathrm{N}(O,G_0^{-1}).$$

4.4 Proof of Theorem 2

If we can show that

$$i_n^{-\frac{1}{2}} l_n \xrightarrow{L} \mathrm{N}(O, K_0), \tag{24}$$

then clearly (Lemma A.3(i))

$$\lim_{M\to\infty} \limsup_{n\to\infty} \mathrm{P}((1/i_n)l'_n l_n > M) = 0,$$

and, since conditions Al-A3 of Theorem 1 are satisfied (because of B1, B2, and B10-B12), both conclusions of Theorem 1 hold. In particular, we obtain

$$\operatorname{plim} i_n^{-\frac{1}{2}}(j_n(\hat{\gamma}_n - \gamma_0) - G_0^{-1} l_n) = 0,$$

which together with (24) implies

$$i_n^{-\frac{1}{2}} j_n(\hat{\gamma}_n - \gamma_0) \xrightarrow{L} \mathrm{N}(O, G_0^{-1} K_0 G_0^{-1}),$$

using Lemma A.1(iv). Hence it suffices to prove (24).

Let $\{x_{n,t}, 1 \leqslant t \leqslant n, n \in \mathbb{N}\}$ be the double sequence of random $p \times 1$ vectors defined by $x_{n,t} := i_n^{-\frac{1}{2}} \xi_t$. Notice that $\xi_n = l_n - l_{n-1}$, so that $l_n = \Sigma_{t=1}^n \xi_t$. To prove (24) we have to show that

$$s_n := \sum_{t=1}^{n} x_{n,t} \xrightarrow{L} \mathrm{N}(O, K_0), \tag{25}$$

which will follow if $\{x_{n,t}\}$ satisfies the conditions of Proposition 1 (section 3).

Now, condition B4 implies that $\{\xi_n\}$ is a vector martingale difference (see Lemma 1; the requirement that the p components of ξ_n are measurable functions of $y_1, \cdots, y_n$ with finite expectation, follows from B3), and hence that $\{x_{n,t}\}$ is a vector martingale difference array. As a consequence of B4 we have

$$\operatorname{cov}(\xi_{ti}, \xi_{sj}) = 0 \quad (t \neq s), \quad \mathrm{E}\xi_{tj} = 0 \quad (t \geqslant 2),$$

so that

$$\begin{aligned}
\mathrm{E} \sum_{t=1}^{n} \xi_{ti} \xi_{tj} &= \sum_{t=1}^{n} \mathrm{E} \xi_{ti} \xi_{tj} \\
&= \sum_{t=1}^{n} \operatorname{cov}(\xi_{ti}, \xi_{tj}) + \sum_{t=1}^{n} (\mathrm{E}\xi_{ti})(\mathrm{E}\xi_{tj}) \\
&= \operatorname{cov}\Big(\sum_{t=1}^{n} \xi_{ti}, \sum_{t=1}^{n} \xi_{tj}\Big) + \Big(\mathrm{E} \sum_{t=1}^{n} \xi_{ti}\Big)\Big(\mathrm{E} \sum_{t=1}^{n} \xi_{tj}\Big) \\
&= \mathrm{E}(\partial \Lambda_n(\gamma_0)/\partial \gamma_i)(\partial \Lambda_n(\gamma_0)/\partial \gamma_j).
\end{aligned} \tag{26}$$

This, in conjunction with B5, implies

$$(1/i_n)\mathrm{E}\max_{1\leq t\leq n}\sum_{j=1}^{p}\xi_{tj}^2\leq(1/i_n)\mathrm{E}\sum_{t=1}^{n}\sum_{j=1}^{p}\xi_{tj}^2=1,$$

so that condition (i) of Proposition 1 holds. Condition (ii) follows from B7. Finally, to prove condition (iii) of Proposition 1, we note, from (26) and condition B6, that

$$\lim_{n\to\infty}(1/i_n)\mathrm{E}\sum_{t=1}^{n}\xi_t\xi'_t=K_0. \tag{27}$$

Letting

$$v_{tij}:=\mathrm{E}(\xi_{ti}\xi_{tj}|y_1,\cdots,y_{t-1}),$$

we obtain

$$\begin{aligned}&(1/i_n^2)\mathrm{var}\sum_{t=1}^{n}\xi_{ti}\xi_{tj}\\&=(1/i_n^2)\mathrm{var}\sum_{t=1}^{n}(v_{tij}+\xi_{ti}\xi_{tj}-v_{tij})\\&\leq(2/i_n^2)\mathrm{var}\sum_{t=1}^{n}v_{tij}+(2/i_n^2)\ \mathrm{var}\sum_{t=1}^{n}(\xi_{ti}\xi_{tj}-v_{tij})\\&=(2/i_n^2)\mathrm{var}\sum_{t=1}^{n}v_{tij}+(2/i_n^2)\sum_{t=1}^{n}\mathrm{var}(\xi_{ti}\xi_{tj}-v_{tij}),\end{aligned}$$

using the fact that $\{\xi_{ti}\xi_{tj}-\mathrm{E}(\xi_{ti}\xi_{tj}|y_1,\cdots,y_{t-1}),t\in\mathbb{N}\}$ is a sequence of uncorrelated random variables. (See Remark 2 in section 4.3.) Conditions B8 and B9 then imply

$$\lim_{n\to\infty}(1/i_n^2)\mathrm{var}\sum_{t=1}^{n}\xi_{ti}\xi_{tj}=0,\quad(i,j=1,\cdots,p). \tag{28}$$

Condition (iii) of Proposition 1 now follows from (27) and (28).

We conclude that $\{x_{n,t}\}$ as defined above satisfies all conditions of Proposition 1 so that (25), and hence (24), holds. This concludes the proof of Theorem 2.

Acknowledgements

We wish to thank P. Robinson and B.B. van der Genugten for their helpful comments, and the Netherlands Organization for the Advancement of Pure Research (Z.W.O.) for financial support. An earlier version of this paper was presented at the 1984 European meeting of the Econometric Society in Madrid.

References

AMEMIYA, T. (1973), Regression analysis when the dependent variable is truncated normal, *Econometrica 41,* 997-1016.

ANDERSON, T.W. (1959), On asymptotic distributions of parameters of stochastic difference equations, *Annals of Mathematical Statistics 30,* 676-687.

APOSTOL, T.M. (1974), *Mathematical analysis,* 2nd edition. Addison-Wesley Reading, Mass.

BAHADUR, R.R. (1967), Rates of convergence of estimates and test statistics, *Annals of Mathematical Statistics 38,* 303-324.

BAR-SHALOM, Y. (1971), On the asymptotic properties of the maximum-likelihood estimate obtained from dependent observations, *Journal of the Royal Statistical Society Series B 33,* 72-77.

BASAWA, I.V., P.D. FEIGIN and C.C. HEYDE (1976), Asymptotic properties of maximum likelihood estimators for stochastic processes, *Sankhyā- Series A 38,* 259-270.

BHAT, B.R. (1974), On the method of maximum likelihood for dependent observations, *Journal of the Royal Statistical Society Series B 36,* 48-53.

BIERENS, H.J. (1981), *Robust methods and asymptotic theory in nonlinear econometrics,* Lecture notes in economics and mathematical systems, Vol. 192, Springer-Verlag. Berlin/Heidelberg/New York.

BILLINGSLEY, P. (1961), *Statistical inference for Markov processes,* University of Chicago Press. Chicago, Illinois.

BRADLEY, R.A. and J.J. GART (1962), The asymptotic properties of *ml* estimators when sampling from associated populations, *Biometrika 49,* 205-214.

CRAMÉR, H. (1946), *Mathematical methods of statistics,* Princeton University Press, Princeton.

CROWDER, M.J. (1976), Maximum likelihood estimation for dependent observations, *Journal of the Royal Statistical Society Series B 38,* 45-53.

DIANANDA, P.H. (1953), Some probability limit theorems with statistical applications, *Proceedings of the Cambridge Philosophical Society 49,* 239-246.

DOOB, J.L. (1953), *Stochastic processes,* John Wiley, New York.

DUGUÉ D. (1937), Application des propriétés de la limite au sens du calcul des probabilités a l'étude de diverses questions d'estimation, *Journal de l'Ecole Polytechnique 3 No. 4,* 305-372.

EICKER, F. (1964), Ueber den zentralen grenzwertsatz fur abhangige zufallsvariable, *Zeitschrift für Wahrscheinlichkeitstheorie 3,* 193-203.

GOURIEROUX, C. and A. MONTFORT (1981), Asymptotic properties of the maximum likelihood estimator in dichotomous logit models, *Journal of Econometrics 17,* 83-97.

HALL, P. and C.C. HEYDE (1980), *Martingale limit theory and its application,* Academic Press New York.

HANNAN, E.J. (1970), *Multiple time series,* John Wiley New York.

HEIJMANS, R.D.H. and J.R. MAGNUS (1986a), Consistent maximum-likelihood estimation with dependent observations: the general (non-normal) case and the normal case, *Journal of Econometrics 32,* to appear.

HEIJMANS, R.D.H. and J.R. MAGNUS (1986b), Asymptotic normality of maximum likelihood estimators obtained from normally distributed but dependent observations, *Econometric Theory 2,* to appear.

HOADLEY, B. (1971), Asymptotic properties of maximum likelihood estimators for the independent not identically distributed case, *Annals of Mathematical Statistics 42,* 1977-1991.

JENNRICH, R.I. (1969), Asymptotic properties of non-linear least squares estimators, *Annals of Mathematical Statistics 40,* 633-643.

MCLEISH, D.L. (1974), Dependent central limit theorems and invariance principles, *Annals of Probability 2,* 620-628.

MORAN, P.A.P. (1971), Maximum-likelihood estimation in non-standard conditions, *Proceedings of the Cambridge Philosophical Society 70,* 441-450.

NORDBERG, L. (1980), Asymptotic normality of maximum likelihood estimators based on independent, unequally distributed observations in exponential family models, *Scandinavian Journal of Statistics 7,* 27-32.

PAGAN, A. (1980), Some identification and estimation results for regression models with stochastically varying coefficients, *Journal of Econometrics 13,* 341-363.

PHILIPPOU, A.N. and G.G. ROUSSAS (1973), Asymptotic normality of the maximum likelihood estimate in the independent not identically distributed case, *Annals of the Institute of Statistical Mathematics 27,* 45-55.

RAO, C.R. (1973), *Linear statistical inference and its applications,* 2nd edition. John Wiley New York.

ROUSSAS, G.G. (1968), Asymptotic normality of the maximum likelihood estimate in Markov processes, *Metrika 14,* 62-70.

ROZANOV, YU. A. (1967), *Stationary random processes,* Holden-Day San Francisco/ Cambridge/ London/ Amsterdam.

SCHÖNFELD, P. (1971), A useful central limit theorem for *m*-dependent variables, *Metrika 17,* 116-128.

SILVEY, S.D. (1961), A note on maximum-likelihood in the case of dependent random variables, *Journal of the Royal Statistical Society Series B 23,* 444-452.

SWEETING, T.J. (1980), Uniform asymptotic normality of the maximum likelihood estimator, *The Annals of Statistics 8,* 1375-1381.

WALD, A. (1948), Asymptotic properties of the maximum likelihood estimate of an unknown parameter of a discrete stochastic process, *Annals of Mathematical Statistics 19,* 40-46.

WEISS, L. (1971), Asymptotic properties of maximum likelihood estimators in some nonstandard cases, *Journal of the American Statistical Association 66,* 345-350.

WEISS, L. (1973), Asymptotic properties of maximum likelihood estimators in some nonstandard cases, II, *Journal of the American Statistical Association 68,* 428-430.

WHITE, H. and I. DOMOWITZ (1984), Nonlinear regression with dependent observations, *Econometrica 52,* 143-161.

WILKS, S.S. (1962), *Mathematical statistics,* John Wiley, New York.

WITTING, H. and G. NÖLLE (1970), *Angewandte mathematische statistik,* B.G. Teubner, Stuttgart.

Received May 1985, revised september 1985.

R.D.H. Heijmans
Institute of Actuarial Sciences
and Econometrics
University of Amsterdam
Jodenbreestraat 23
1011 NH Amsterdam
The Netherlands

J.R. Magnus
London School of Economics
and Political Science
Department of Economics
Houghton Street
London WC2A 2AE
United Kingdom

Appendix

In this appendix, $\{x_n\}$ and $\{y_n\}$ are two sequences of random $p \times 1$ vectors, $\{A_n\}$ is a sequence of random $p \times p$ matrices, and $\{k_n\}$ is a sequence of nonrandom real numbers. Below we state, without proof, some simple results which are used in the main text.

Lemma A.1

(i) $\operatorname{plim} A_n = A, \operatorname{plim} x_n = x \Rightarrow \operatorname{plim} A_n x_n = Ax$;

(ii) $\operatorname{plim} A_n = I_p, \operatorname{plim} A_n x_n = O \Rightarrow \operatorname{plim} x_n = O$;

(iii) $\operatorname{plim} x_n = O$, $\{k_n\}$ is bounded $\Rightarrow \operatorname{plim} k_n x_n = O$;

(iv) $\operatorname{plim}(x_n - y_n) = O, x_n \xrightarrow{L} x \Rightarrow y_n \xrightarrow{L} x$.

Lemma A.2.

Assume that $\operatorname{plim} A_n = O$, and

$$\lim_{M \to \infty} \limsup_{n \to \infty} P(x'_n x_n > M) = O. \tag{A.1}$$

Then, $\operatorname{plim} A_n x_n = O$.

Lemma A.3

Any of the following two conditions is sufficient for (A.1):

(i) $x_n \xrightarrow{L} x$;

(ii) $E(x'_n x_n)^r$ is bounded in n for some $r > 0$.